Learning to Label Sequences in One Pass

Antoine Bordes
NEC Laboratories America, Inc.
LIP6, Université de Paris 6,
<antoine.bordes@lip6.fr>

Nicolas Usunier
LIP6, Université de Paris 6,
<nicolas.usunier@lip6.fr>

Léon Bottou
NEC Laboratories America, Inc.
<leonb@nec-labs.com>

The sequence labelling task consists in predicting a sequence of labels given an observed sequence of *tokens*. This task is an example of *structured output* learning system and appears in practical problems in computational linguistics and signal processing.

Two informal assumptions are crucial for this task. The first states that a label depends only on the surrounding labels and tokens. The second states that this dependency its invariant with is time index. These assumptions are expressed through the parametric formulation of the models, and, in the case of probabilistic models, through conditional independence assumptions (Markov models). Part of the model specification is then the *inference procedure* that recovers the predicted labels for any input sequence.

Batch sequence learning algorithms determine the model parameters by optimizing a global objective function that depends on all the training sequences. This approach is compatible with a variety of inference procedures. However the computational cost of learning usually grows faster than the total number of tokens in the training set.

Online sequence learning algorithms are less costly because they iteratively update the model parameters by separately processing each training sequence, or each training token. Although algorithm of the latter kind are restricted to models based on *greedy inference*, they have been shown to be extremely competitive in practice.

Following [2], we cast both exact and greedy inference as two quadratic programming problems whose kernel matrices define the same feature space and then derive *two online sequence learning algorithms* using a slightly simplified (improved) variant of the LaRank algorithm [1]. Both algorithms *empirically perform as well as the equivalent batch algorithm with exact inference* with only one epoch over the training data. Their training times *scale linearly with the number of training tokens*. Since both algorithms derive from the same setup we can also discuss the observed differences in training time and sparsity that tend to favor the greedy online algorithm.

References

[1] Bordes, A. & Bottou, L. & Gallinari, P. & Weston, J. (2007) Solving multiclass support vector machines with LaRank, *Proceedings of the 24th International Conference on Machine Learning (ICML07)*.

[2] Tsochantaridis, I. & Joachims, T. & Hofmann, T. & Altun, Y. (2005) Support vector machines for structured and interdependent output variables, *Journal of Machine Learning Research, 6*, 1453-1484.

Topics: Online Learning, Structured Prediction.

Preference: Poster.